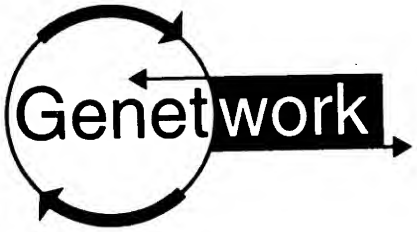

Go hunting in sequence databases but watch out for the traps

The large amount of data created by world-wide sequencing efforts calls for automation in data handling and analysis. This requires accurate storage and updating mechanisms as well as appropriate retrieval software. User-friendly interfaces are also needed, as the number of researchers that access the information stored in public sequence databases is increasing considerably. Although the database teams are aware of the demands and the invaluable sequence databases are improving, they are also the product of history and, like the accessing software, far from perfect. Thus, at present, working with sequence databases requires knowledge about their powers and their pitfalls. Here, we concentrate on some of the problems that many users are unaware of, but that can have a considerable influence on the interpretation of the data. Some of the more frequent problems are summarized below, and some specific examples are given in Boxes 1 and 2.

Problems within the sequence databases themselves

Sequencing errors seem to be in the order of 0.1% (Ref. 1) (excluding ESTs, single reads with a very high error rate) affecting about 5% of the proteins[2]. When screening 300 human proteins in SWISS-PROT that have been published separately more than once, we find that 0.3% of the amino acids are different; this is a lower limit as lots of corrections have already been done and as sequences appearing in two different publications are often not independent. In any case, only frameshift errors and artificial stop codons can be detected unambiguously; point mutations are hard to verify as natural polymorphism or strain differences cannot easily be excluded. Even if this rate seems low, errors can accumulate in the sequence of interest (see Ref. 3 for an example) and can lead to functional misinterpretations. Moreover, although the quality of sequencing is improving, budget calculations might favor quantity instead of quality in the near future; the successful strategies based on ESTs demonstrate that data quality and its interpretation remain a major issue. Errors of various sources are also a major problem for other molecular databases such as Brookhaven Protein Data Bank[4].

The processing of raw DNA by database management software is another serious source of problems. For example, false translation of genomic DNA into gene products, having missed exons or translated introns, leads to erroneous entries in protein sequence databases; the correct initiating methionine is not always chosen as a translation start; or ORFs translated from the opposite strand of the gene end up as proteins. The challenge is to improve prediction methods as the widely used algorithms of gene identification in higher eukaryotes have only an accuracy between 60–70% (Ref. 5): nearly a third of the automatically predicted proteins from genomic DNA without clear homologs are expected to contain some serious errors.

Erroneous annotation is also common, ranging from simple spelling errors to

Box 1. Some arbitrarily chosen examples that demonstrate various kinds of pitfalls in database usage

Synonyms

In organisms that are the target of major genetic studies, it often happens that the same gene is isolated by many different groups and so it ends up with many different names. For example, yeast *TUP1* is also known as *AER2*, *SFL2*, *CYC9*, *UMR7*, *AAR1*, *AMM1* and *FLK1*. In *Escherichia coli*, *hns* is also known as *hnsA*, *drdX*, *osmZ*, *bglY*, *msyA*, *cur*, *pilG* and *topS*. The multiplicity of synonyms also exists at the level of protein names. For example, annexin V was also called: lipocortin V, endonexin II, calphobindin I, placental anticoagulant protein I, pp4, thromboplastin inhibitor, vascular anticoagulant-alpha and anchorin CII.

Different gene – same name

Conversely, it often happens that the same gene name is given to two different genes. Generally one of these duplicate names is quickly changed, but in some cases the two gene names each find a lobby and are simultaneously promoted. For example, yeast *MRF1* is both the gene for the mitochondrial peptide chain release factor 1 and for the mitochondrial respiratory function protein 1. A famous example is 'cyclin', the accepted name for a large family of cell-cycle components, which became so prominent that this name is no longer used for a protein now known as proliferating cell nuclear antigen (originally called cyclin).

Spelling

Even spelling mistakes can end up as gene synonyms. For example, the yeast gene, *SCD25* (suppressor of *CDC25*), was so often misquoted as *SDC25* that it has become an accepted synonym. In addition to spelling mistakes, database queries can be hindered by: differences between US and UK spelling (e.g. hemoglobin or haemoglobin); representation of special characters, such as accented characters (e.g. *Krüppel*, *Krueppel* or *Kruppel*); upper and lower cases (e.g. in the *Drosophila* genetic nomenclature, *H* is *Hairless*, but *h* is *hairy*).

Biological source and contamination

There are numerous problems with the annotation of the biological source of a sequence. For example, the ORGanelle division of EMBL/GenBank division should only contain sequences that are encoded on the mitochondria or plastid. But often entries reporting nuclear-encoded genes for proteins targeted to such an organelle are wrongly entered in the ORG division. The converse is also true as some chloroplast or mitochondrial encoded sequences are sometimes found in other divisions of EMBL/GenBank. This problem can have an effect on the derived protein sequence: if a nuclear-encoded mitochondrial gene is misclassified into the ORG section, the resultant translation will be wrong as the automatic translation software will assume that a mitochondrial genetic code should be used. The contamination of cDNA libraries (usually by fungal or bacterial DNA) is still an issue (a prominent example is one of Genethon's 'human' EST libraries that have a surprising number of matches with the yeast genome). Some scientific surprises can result from these issues: it was found recently (Laurent Duret, pers. commun.) that the sequence of two genes coding for annexin 1 and insulin from a sponge (their 'identification' in lower eukaryotes was unexpected) were too closely related to their mammalian homologs. It turned out that the biological samples had most probably been contaminated by an undetermined rodent species.

discrepancies between sequences published in printed form. This should decrease with electronic submissions and quality check software. Sequences can also be incorrectly labelled because of contamination of the sequenced material (Box 1).

A major annotation problem is the functional description of a gene. Due to the increasing number of 'software robots'[6] that automatically assign functions based on similarities, a single wrongly annotated entry will lead to whole families with artificial functions based on similarities to that entry [e.g. Ref. 7 in which information was carefully transferred from a gene called *nifr3* to several unannotated proteins; whether *nifr3* functions, however, in nitrogen fixation (nif) remains unclear, despite its annotation]. Automatic methods often cannot adapt to the dynamic annotation of entries and might, for example, point to neighboring genes in automatically translated protein sequences after a new ORF has been discovered in an already annotated region. Further annotation problems arise due to user interpretation (see below).

The retrieval of data is often hindered by incorrect genetic nomenclature (Box 1). Even in relatively similar organisms, such as budding yeast and fission yeast, the same gene name actually points to non-homologous functionally different proteins (e.g. *RAD4*). Details in the syntax can often not be reflected in the databases. Major attempts have been made for classifying enzymes, but even here the functional classification needs to be complemented by consistent homology and three-dimensional information. Nomenclature is also needed at the level of domains, structurally independent parts of proteins for which frequently several names coexist in the literature and consequently in the databases[8]. The different communities should force standardization.

Finally, databases are not always up to date regarding the functional information or other annotated features because there is currently no systematic update mechanism. Due to the policy of some databases that only authors can change the content of an entry, followup characterizations of genes or gene products are, thus, only occasionally included.

Problems of interpretation

Numerous pitfalls are related to the interpretation of the results of the database-accessing software; simple problems arise if the retrieval system does not access the full dataset so that stored information is not found. Furthermore, the occasional user often only accesses the major sequence databases that contain information from all organisms. Many communities studying particular protein families or organisms know about specialized databases that contain much more information on particular genes or proteins, but which is often not linked to the major databases. There is hope that this will change in the near future (e.g. links to FlyBase, YPD in SWISS-PROT).

A battery of traps result from database similarity searches, probably the most prominent form of database access. For example, the user might have insufficient knowledge of the limits of the programs (e.g. 'homology' to the coiled-coil region of myosin that is due to similar structural constraints) and inadequate thresholds and parameters often prevent an objective analysis. A different problem results from the pressure on sequencing groups, not to overlook interesting functional information in 'their' sequences. Thus, similarity search methods are stretched and spurious hits are taken as real. Moreover, similarities might only be restricted to certain domains, but the function is transferred to a whole protein. All such questionable interpretations end up in databases and are then considered as facts.

Finally, here is just one example that demonstrates the difficulties of functional predictions based on homology. Imagine the best hit to your *Drosophila* sequence is the human zinc-containing alcohol dehydrogenase class 4 μ/σ (in databases mu/sigma) chain (ADH7). It is very difficult to find out, whether the *Drosophila* sequence is the ortholog, another alcohol dehydrogenase, a homologous lactate dehydrogenase, a more distantly related oxidoreductase, or perhaps just a protein with an NADH-binding site. There is currently little quantification possible, in terms of functional similarity; a way out might be the knowledge of the complete

Box 2. Unusual database entries

'Protein' sequences in databases can be as short as one amino acid that is sometimes an X (as happens in the patent divisions of the databases) so that several database accession software packages have problems. Automatic DNA translation programs that contribute a considerable fraction of the protein sequences can also be mislead: the following TREMBL (automatic Translation of EMBL; version August 96) entry is a mistranslation (compare with the annotated CDS), probably due to some annotation problems in the corresponding EMBL entry. The name is also unusual: a human protein with an identifier starting with MM (usually meaning *Mus musculus*). It is supposed to encode a small region of trk4 but the translation comes up with parts of a different protein, MAC25. Detective work is needed to figure out the errors that lead to the wrong translation.

```
ID   MMTRK4A_1  standard; PRT; 118 AA.
AC   M55337;
DR   EMBL; M55337; MMTRK4A.
DE   gene: "trk4"; product: "oncogene tyrosine protein kinase receptor";
DE   Human oncogene tyrosine protein kinase receptor (trk4) mRNA,
DE   partial cds.
OS   Homo sapiens (human)
OC   Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
OC   Theria; Eutheria; Primates; Haplorhini; Catarrhini; Hominidae.
FT   CDS             <1. .>354
FT                   /gene="trk4"
FT                   /note="NCBI gi:. 339916"
FT                   /codon_start=1
FT                   /product="oncogene tyrosine protein kinase receptor"
FT                   /db_xref="PID:g339916"
FT                   /translation="HSIKDVHARLQALAQEQEEEEQEQEEQEGEEAATPSGGGRNRSAS
FT                   SSWVGTMAGISMSLHFMTLGGGSSLSPTEGKGSGLQGHIIENPQYFSDACVHHIKRRDIV
FT                   LKWELGEGAFGKVF"
CC   translated using genetic code table "Standard"
CC   Warning: codon start shifted by 1
CC   Warning: illegal start codon
SQ   Sequence   118 BP;

Mmtrk4a_1  Length: 118  August 20, 1996 23:39  Type: P  Check: 7282  ..
       1  PLPPHPAMER PSLRALLLGA AGLLLLLLPL SSSSSSDTCG PCEPASCPPL
      51  PPLGCLLGET RDACGCCPMC ARGEGEPCGG GGAGRGYCAP GMECVKSRKR
     101  RKGKAGAAAG GPGVSGVC
```

gene pool of organisms from all major taxa, which will allow classification within multigene families via phylogenetic trees.

Shared responsibilities

Although the list of problematic issues is much longer, we wish to point out that sequence databases are the most useful tool in sequence analysis and the question should be how can one further improve their value by enhancing the data storage, handling and retrieval? How should the responsibility for this task be shared? Everybody who stores information should feel responsible for the data and the annotation quality. Database teams have a restricted budget and can only provide some quality checks (e.g. for cloning and sequencing errors, artificially translated vectors, repeats and so on). Databases rely on standards and these have also to come from the different communities in the form of agreed nomenclature and clearly reproducible functional characterizations. Specialists should spend the time to give feedback on encountered problems and database teams should have mechanisms to include such improvements. This is, of course, easily said, but opinions about data and annotation vary and the truth is not always obvious. In conclusion, a concerted effort is needed from the database teams that have to maximize their services and the user community that should share responsibilities in taking care of the quality of the entries.

Peer Bork
bork@embl-heidelberg.de

EMBL, Meyerhofstr. 1, 69012 Heidelberg and Max-Delbrück-Center for Molecular Medicine, Berlin-Buch, Germany.

Amos Bairoch
bairoch@cmu.unige.ch

Medical Biochemistry Department, University of Geneva, 1211 Geneva 4, Switzerland.

References

1 States, D.J. (1992) *Trends Genet.* 8, 52–55
2 Birney, E., Thompson, J. and Gibson, T. *Nucleic Acids Res.* (in press)
3 Bork, P. (1996) *Science* 271, 1431–1432
4 Hooft, R.W.W. *et al.* (1996) *Nature* 381, 272
5 Burset, M. and Guigo, R. (1996) *Genomics* 34, 353–367
6 Boguski, M. and McEntyre, J. (1994) *Trends Biochem. Sci.* 19, 71
7 Casari, G. *et al.* (1995) *Nature* 376, 647–648
8 Bork, P. and Bairoch, A. (1995) *Trends Biochem. Sci.* 20 (Suppl. C03) Poster

Genetwork is a regular column of news and information about Internet resources for researchers in genetics and development (pp. 425–427). Genetwork is compiled and edited with the help of Steven E. Brenner (MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK CB2 2QH) and Jeremy Rashbass (Department of Histopathology, Addenbrooke's Hospital, Hills Road, Cambridge, UK CB2 2QQ).

If you would like to announce or publicize an Internet resource, please contact: TIG@elsevier.co.uk

MEETING REPORTS

Flies in Crete

10TH EMBO WORKSHOP ON THE **MOLECULAR AND DEVELOPMENTAL BIOLOGY OF *DROSOPHILA***, KOLYMBARI, CRETE, 14–20 JULY 1996.

The results presented at this meeting were enormously varied and informative and the comments below represent just a small sample of the interesting science that was presented.

The importance of polarity within a single cell was illustrated several times. Transcripts of the segment polarity gene *wingless* (*wg*) were found to be apically localized in polarized epithelial cells (H. Krause, Toronto) and, interestingly, localization was required for *wg* function. In contrast, localized *wg* transcripts were not required for *wg* function in nonpolar mesenchymal cells. Also, with regard to cellular polarity, the *inscuteable* gene was shown to be required for the orientation of the mitotic spindle and, therefore, for the correct plane of cell division (W. Chia, Singapore). Remarkably, in neuroblasts INSCUTEABLE is apically localized and is required for the basal localization of the homeodomain protein, PROSPERO. Thus, INSCUTEABLE appears to be an important component of the positional information within a cell.

Five major signaling pathways were discussed: *hedgehog* (*hh*), *wingless* (*wg*), *decapentaplegic* (*dpp*), EGF and FGF. Perhaps not surprisingly, several intersections and similarites between signaling pathways were apparent. For example, SMOOTHENED protein, which appears to be a G-coupled seven-transmembrane receptor, was suggested to be an HH receptor (M. Noll, Zürich). Curiously, SMOOTHENED shares striking similarity to the FRIZZLED family of proteins, which are putative receptors for Wnt (e.g. WG) signals. Downstream of the WG signal might be HMG-domain transcription factors related to mouse lymphoid enhancer-binding factor 1 (LEF1) (M. Bienz, Cambridge, UK, in collaboration with R. Grosschedl). LEF1 binds to ARMADILLO, another downstream component of the WG signal and, when expressed in *Drosophila*, phenocopies *wg*-overexpression phenotypes. Therefore, the tantalizing hypothesis that LEF1 might be an ARMADILLO-activated nuclear target for WG signaling was suggested. Another molecule downstream of the Wg signal is encoded by *arrow*. Surprisingly, *arrow* turns out to be identical to the gene *centrosomin* which is a component of the centrosome (T. Kaufman, Bloomington and S. DiNardo, New York). How the product(s) of a single gene can participate in two seemingly very different cellular processes provides a curious puzzle. In the embryonic endoderm, the homeodomain protein encoded by *extradenticle*, which binds to DNA cooperatively with HOX proteins, was shown to translocate from cytoplasm to nucleus in response to both DPP and WG signals (R. Mann, New York). Similarly, the novel protein encoded by *Mothers against dpp* was also shown to translocate from the cytoplasm into nuclei in response to DPP (W. Gelbart, Cambridge, USA). Thus, controlling protein localization within a cell might be a common response to extracellular signals.

In addition to intersecting signaling pathways, combinations of different signals were shown to be important for the activation of *even-skipped* expression in a small cluster of mesoderm cells (A. Michelson, Cambridge, USA). The selection of these mesoderm cells, which are founders for a subset of muscles, depends on intersecting fields of *wg* and *dpp* expressing cells in the ectoderm, together with a RAS-dependent pathway (perhaps the EGF pathway). These signals define an equivalence group, which, as is the case for neuroblast selection, is refined by the action of another set of signaling molecules encoded by the neurogenic genes, *Notch* and *Delta*.